A-Tune算法解析ceph调优实践

麒麟软件有限公司　滕磊

公司介绍

KYLINSOFT 麒麟软件

连续十一年，市场第一

历史沿革

2019年12月由天津麒麟和中标软件强强整合

股权架构

国有控股、市场化运营的高新技术企业

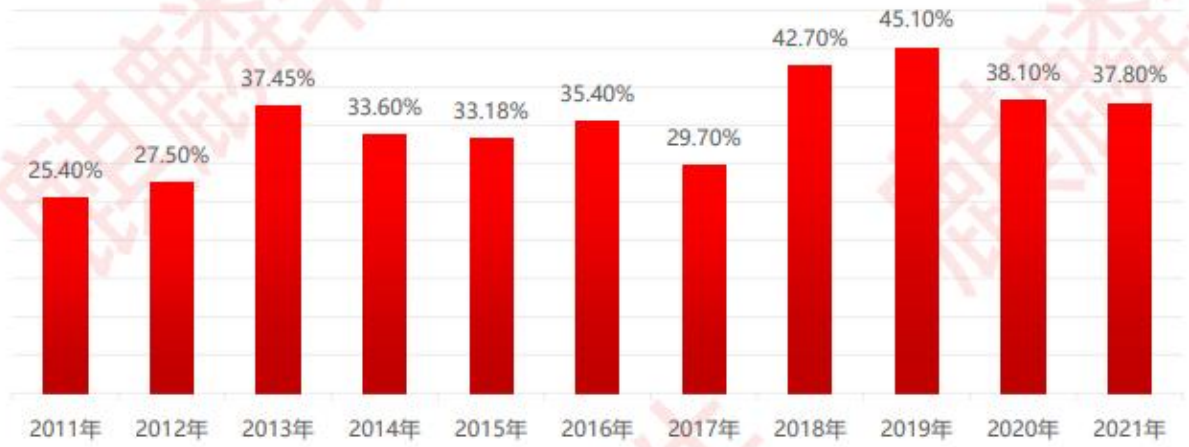
麒麟软件

主营业务

安全可信操作系统的研发与服务

旗下品牌

银河麒麟、中标麒麟、星光麒麟



赛迪顾问统计，麒麟软件连续十一年市场占有率NO.1

银河麒麟V10在桌面政府市场销售额占比第一，占有率超70%

银河麒麟V10在服务器政府市场销售额占比第一，占有率超90%

目录/CONTENTS

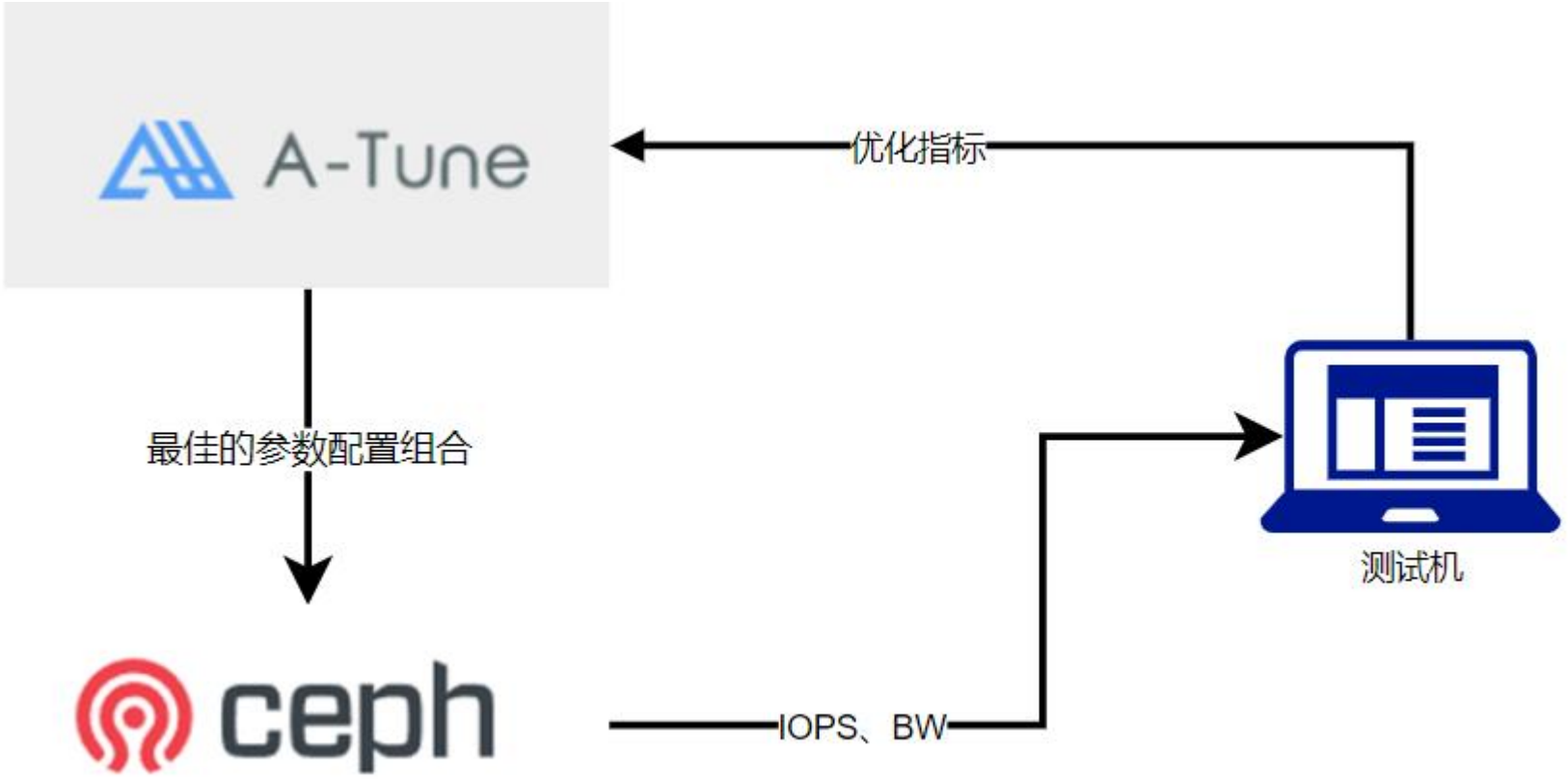
KYLINSOFT 麒麟软件

OpenEuler

01

A-Tune说明

A-Tune说明

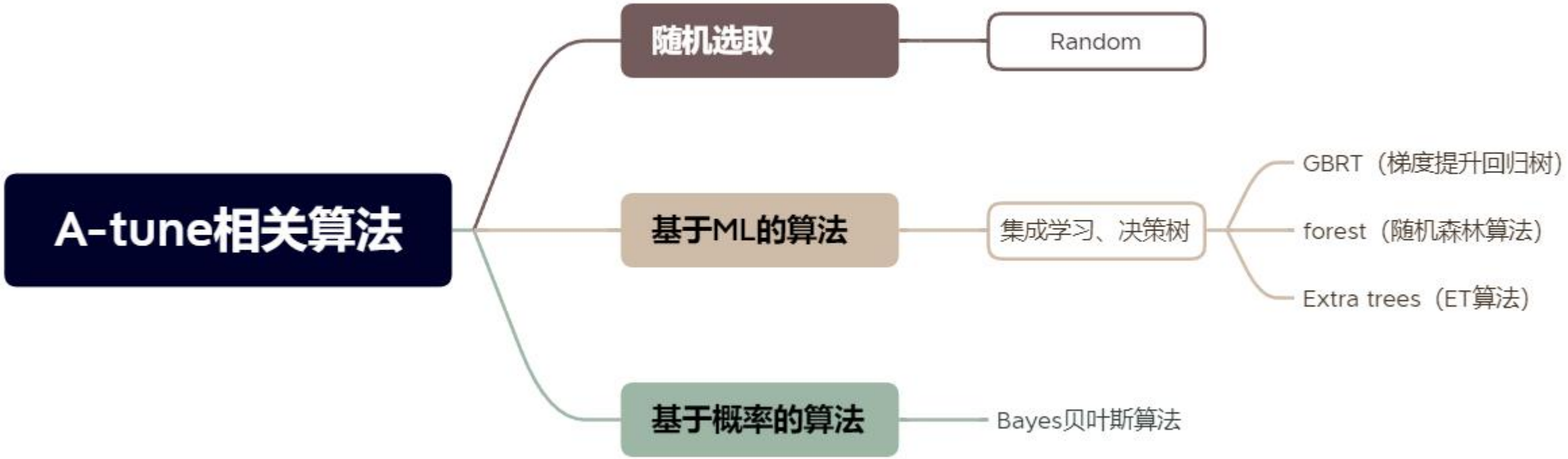


A-Tune是一款基于openEuler开发的，**自动化、智能化**性能调优引擎。它利用人工智能技术，根据业务负载情况动态调节并给出**最佳的参数配置组合**，从而使业务处于最佳运行状态。

02

A-Tune算法介绍

A-Tune相关算法


A-tune相关算法
随机选取
Random
基于ML的算法
集成学习、决策树
GBRT（梯度提升回归树）
forest（随机森林算法）
Extra trees（ET算法）
基于概率的算法
Bayes贝叶斯算法

集成学习


集成学习
概念
通常一个集成学习器的分类性能会好于单个分类器,将多个分类方法聚集在一起,以提高分类的准确率。
思想
好比人做出一个决策时，会从不同方面，不同角度，不同层次去思考，这就会有多个决策结果产生，最终我们会选一种决策作为最终决策。
基学习器
组成集成学习的单个分类器称为基学习器，比如决策树、神经网络单元等
集成方式
同构集成：相同类型的基学习器组成的方法
异构集成：相同类型的基学习器组成的方法
学习模式
串行：个体学习器之间存在强依赖关系，必须串行生成的序列化方法
并行：个体学习器不存在强依赖关系，可以同时生成的并行化方法

GBRT算法

GBRT算法

GBRT是回归树，不是分类树。其核心就在于，每一棵树是从之前所有树的残差中来学习的。

是一种**串行的、同构的**集成学习方法。

优势：

1、模型的准确度较高

2、适应不同类型的数据

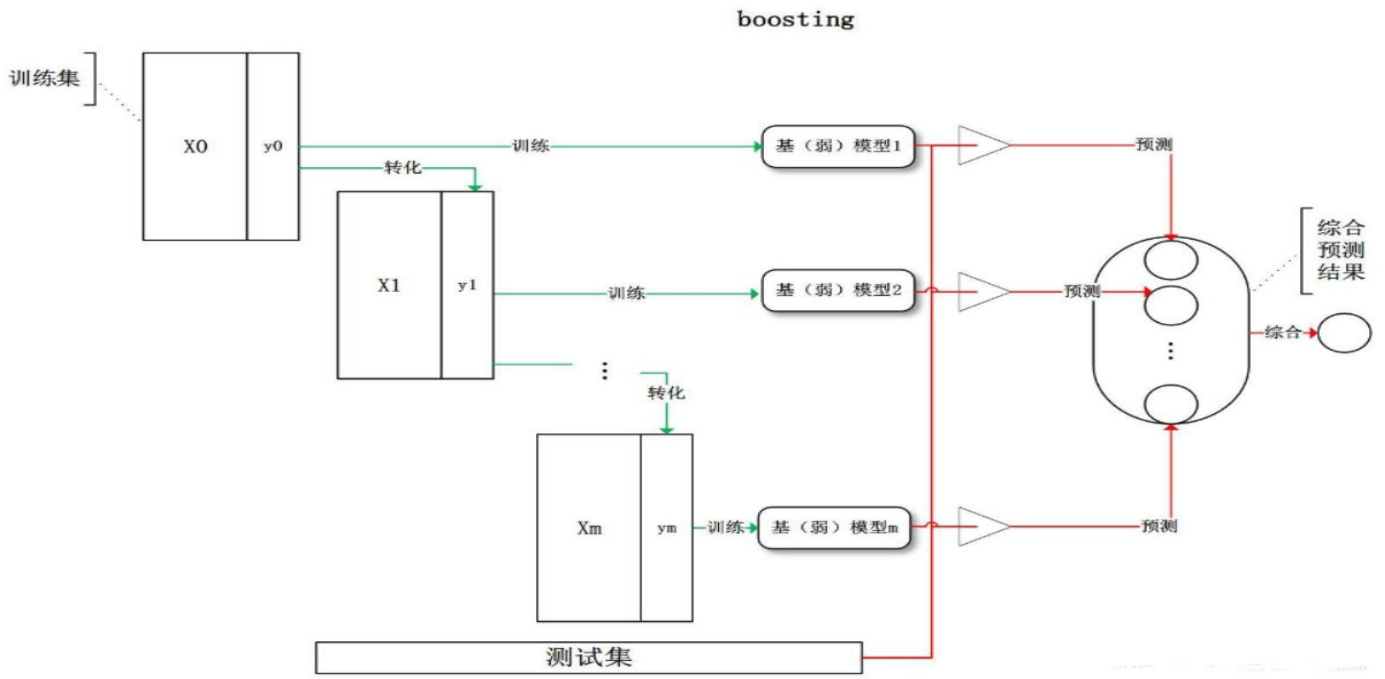
3、对复杂数据和高维数据处理能力更强

缺点：

1、训练时间较长，尤其是在数据量较大

2、可解释性较差

3、对计算资源消耗很大



随机森林算法

随机森林算法

随机森林(Random Forests)是通过多个决策树集成的一种算法，是一种基于Bagging的集成学习方法。是一种**并行的、同构的**集成学习方法。

优势：

1、可解释性较好

2、对特征工程的依赖性较低

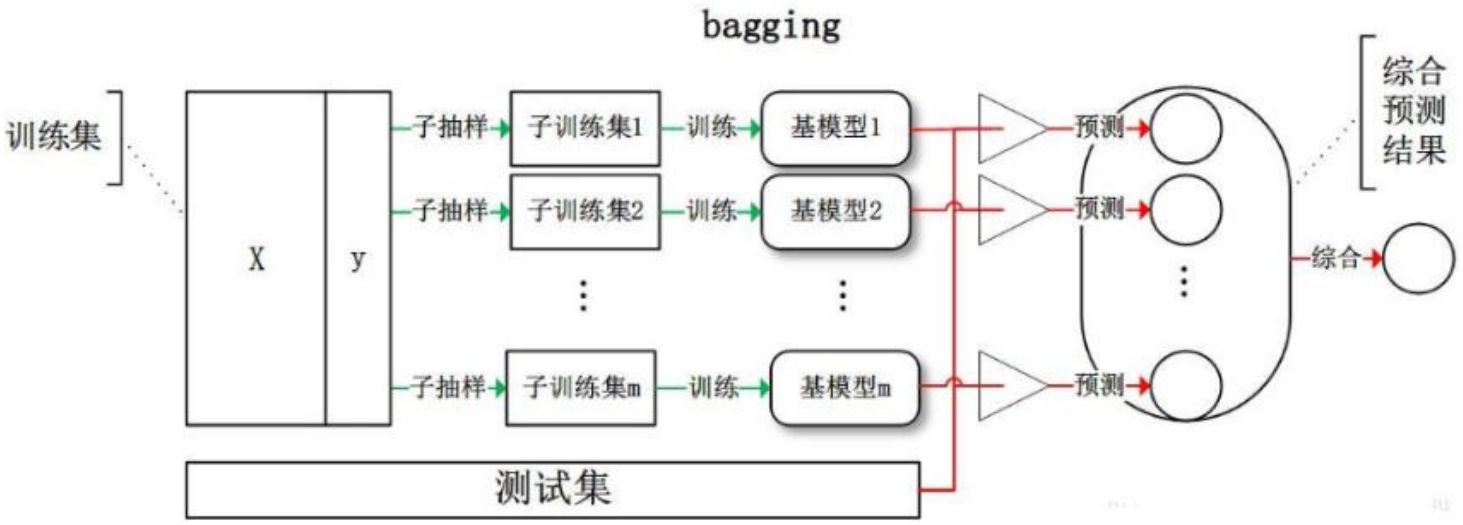
3、对于异常值和噪声的鲁棒性更好

缺点：

1、计算复杂度较高，训练时间较长

2、对高维稀疏数据不够友好

3、对特定类型的数据不兼容



ET算法 & 贝叶斯算法

KYLINSOFT 麒麟软件 OpenEuler

ET算法

ET算法是随机森林算法的一个变种，在随机性程度、选取最优划分点的方式、集成方式等方面存在一些差异；

1、Extra Trees 会使用所有的样本，这样可以减少方差；

2、Extra-Trees 相比于随机森林，拥有更快的训练速度；

贝叶斯

贝叶斯是一种基于概率的算法，对待预测样本进行预测，过程简单。

1、需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种；

2、通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率

KYLINSOFT 麒麟软件 OpenEuler

03

基于A-tune的ceph优化实践

环境搭建

1. 环境信息

硬件信息：　Kunpeng-920虚拟机 8核 8G HDD

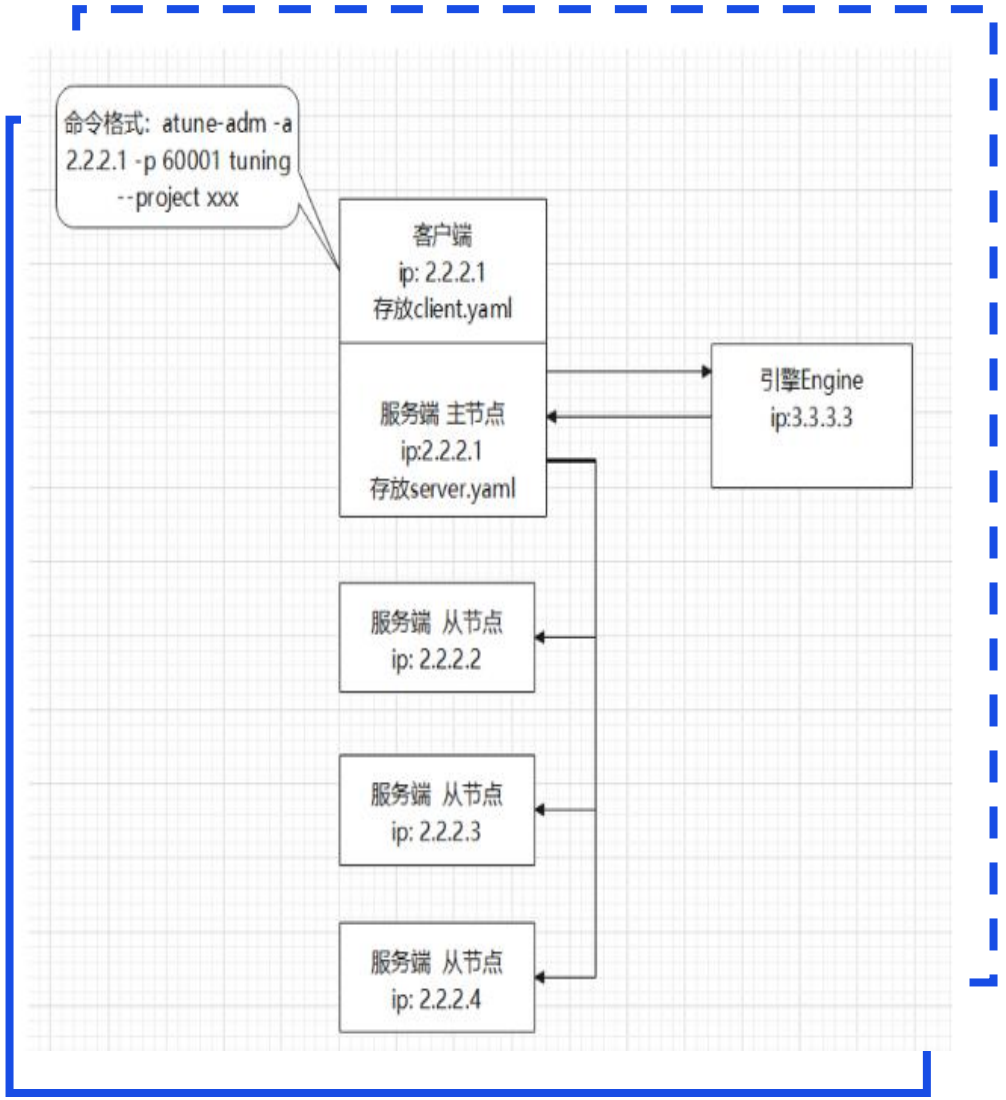
OS信息：　Kylin Linux Advanced Server release V10 SPx

Ceph版本：　14.2.8

Ceph配置：　4台主机，每台主机三块osd

测试工具：　fio-3.7-4

测试指标：　4k顺序读写时的IOPS



2. 集群信息

A-tune集群部署：为了支持多节点场景快速调优，A-Tune支持对多个节点里的参数配置同时进行动态调优，避免用户单独多次对每个节点进行调优，从而提升调优效率。

客户端：　atune、atune-client

引擎：　atune、atune-engine

服务端：　atune

A-Tune配置

```
# ********************************** server ********************************
# atuned config
[server]
# the protocol grpc server running on
# ranges: unix or tcp
protocol                = tcp
# the address that the grpc server to bind to
# default is unix socket /var/run/atuned/atuned.sock
# ranges: /var/run/atuned/atuned.sock or ip address
address                 = 2.2.2.1
# the atune nodes in cluster mode, separated by commas
# it is valid when protocol is tcp
connect = 2.2.2.1,2.2.2.2,2.2.2.3,2.2.2.4
# the atuned grpc listening port
# the port can be set between 0 to 65535 which not be used
port = 60001
# the rest service listening port, default is 8383
# the port can be set between 0 to 65535 which not be used
rest_host               = localhost
rest_port               = 8383
# the tuning optimizer host and port, start by engine.service
# if engine_host is same as rest_host, two ports cannot be same
# the port can be set between 0 to 65535 which not be used
engine_host             = 3.3.3.3
engine_port             = 3838
# when run analysis command, the numbers of collected data.
# default is 20
sample_num              = 20
# interval for collecting data, default is 5s
interval                = 5
# enable gRPC authentication SSL/TLS
# default is false
# grpc_tls = false
# tlsservercafile = /etc/atuned/grpc_certs/ca.crt
# tlsservercertfile = /etc/atuned/grpc_certs/server.crt
# tlsserverkeyfile = /etc/atuned/grpc_certs/server.key
# enable rest server authentication SSL/TLS
# default is true
rest_tls                = false
tlsrestcacertfile       = /etc/atuned/rest_certs/ca.crt
tlsrestservercertfile   = /etc/atuned/rest_certs/server.crt
tlsrestserverkeyfile    = /etc/atuned/rest_certs/server.key
# enable engine server authentication SSL/TLS
# default is true
engine_tls              = false
```

3. 配置A-Tune

修改配置文件：/etc/atuned/atuned.cnf

1.protocol 值设置为tcp

2.address设置为当前节点的ip地址

3.connect设置为所有节点的ip地址，第一个为主节点，其余为从节点ip，中间用逗号隔开。

4.port填写为60001

5.engine_host和engine_port填写为引擎ip和引擎ip的端口号。

6.在调试时，可以设置rest_tls 和engine_tls 为false 。

配置A-Tune引擎

```
################################## engine ##################################
[server]
# the tuning optimizer host and port, start by engine.service
# if engine_host is same as rest_host, two ports cannot be same
# the port can be set between 0 to 65535 which not be used
engine_host = 3.3.3.3
engine_port = 3838

# enable engine server authentication SSL/TLS
# default is true
engine_tls = false
tlsenginecacertfile = /etc/atuned/engine_certs/ca.crt
tlsengineservercertfile = /etc/atuned/engine_certs/server.crt
tlsengineserverkeyfile = /etc/atuned/engine_certs/server.key
```

4. 修改引擎配置文件

修改配置文件：engine.cnf

需要重启服务，配置生效

服务端节点输入命令：systemctl restart atuned

引擎端节点输入命令：systemctl restart atune-engine

配置客户端配置文件

```
project : "ceph_fio"
engine : "gbrt"
iterations : 20
random_starts : 10


benchmark : "fio read-wirte-4k.fio"
evaluations :
        -
         name : "iops"
         info :
                get : "echo '$out'|grep -w iops |grep avg |awk -F '[=,]' '{print $6}'"
                type : "negative"
                weight : 100
```

5. 修改客户端配置文件

project：　　工程名

engine：　　调优算法（forest、gbrt、bayes、extraTrees）

iterations：　　调优迭代次数

get：　　获取性能评估结果的脚本

weight：　　该指标的权重百分比，0-100

type：　　positive代表最小化性能值，negative代表最大化性能值

配置服务端配置文件

```
project: "ceph_fio"
maxiterations: 2048
startworkload: "systemctl start ceph.target"
stopworkload: "systemctl stop cceph.target"
object :
  -
    name : "rbd_cache_target_dirty"
    info :
        desc : ""
        get : "cat /etc/ceph/ceph.conf | grep '^rbd_cache_target_dirty =' | awk '{ $1=NULL; print $3 }'"
        set : "sed -i 's/^rbd_cache_target_dirty.*$/rbd_cache_target_dirty = $value/g' /etc/ceph/ceph.conf"
        needrestart: "true"
        type : "continuous"
        scope :
          - 16777216
          - 268435456
        items :
        dtype : "int"
```

6. 修改服务端配置文件

startworkload/stopworkload： 启动/停止服务的脚本

get： 获取参数脚本

set： 设置参数脚本

scope： 最大值和最小值的选取

needrestart： 是否需要重启

KYLINSOFT 麒麟软件 OpenEuler

04

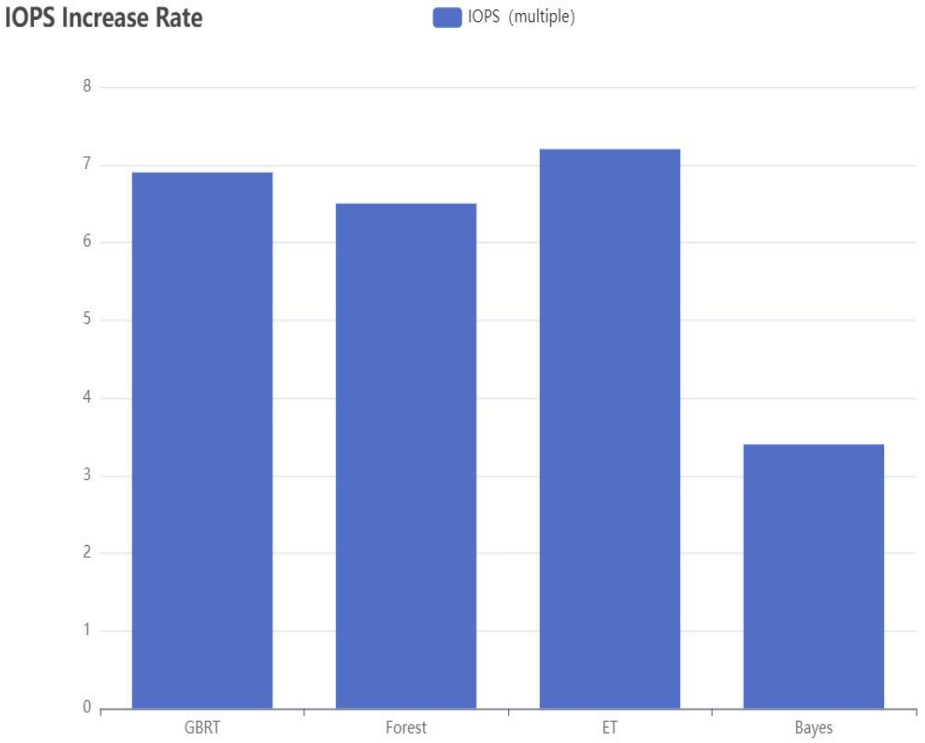
实验结果分析

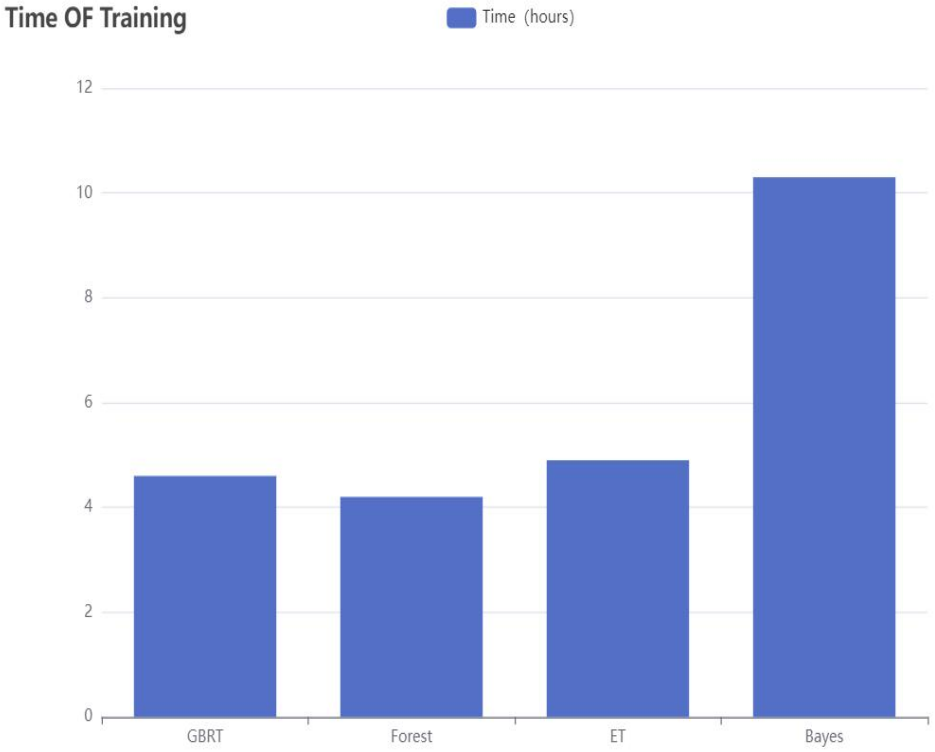
实验结果分析

4K顺序写

iops提升倍数：GBRT（6.9倍）、forest（6.5倍）、ET（7.2倍）、Bayes（3.4倍）

训练时长：GBRT（4.5小时）、forest（3.9小时）、ET（5.2小时）、Bayes（10.4小时）



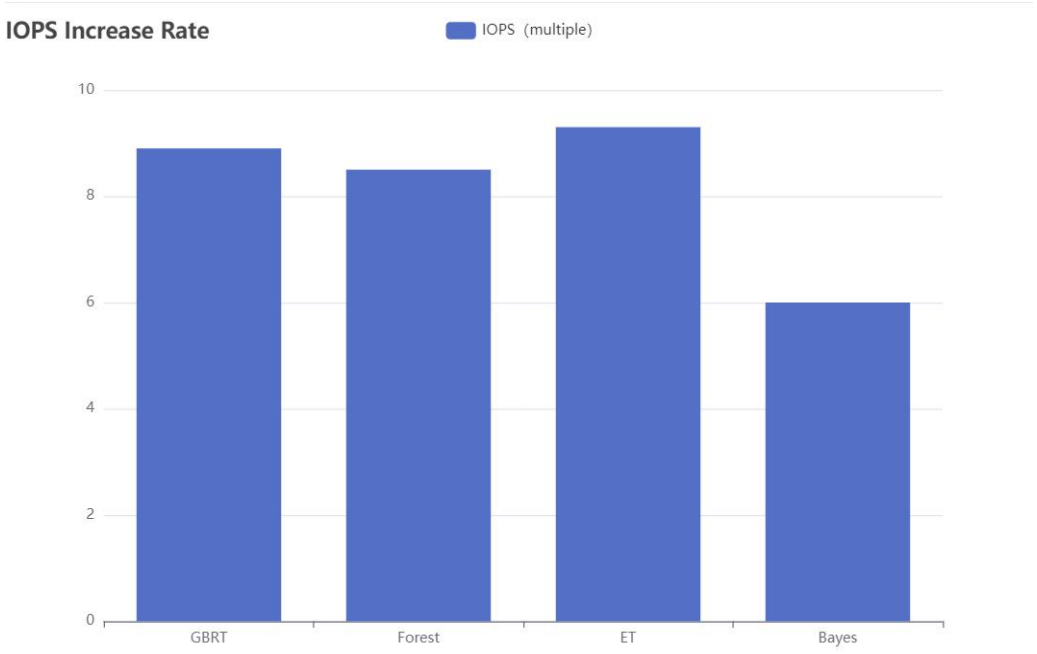


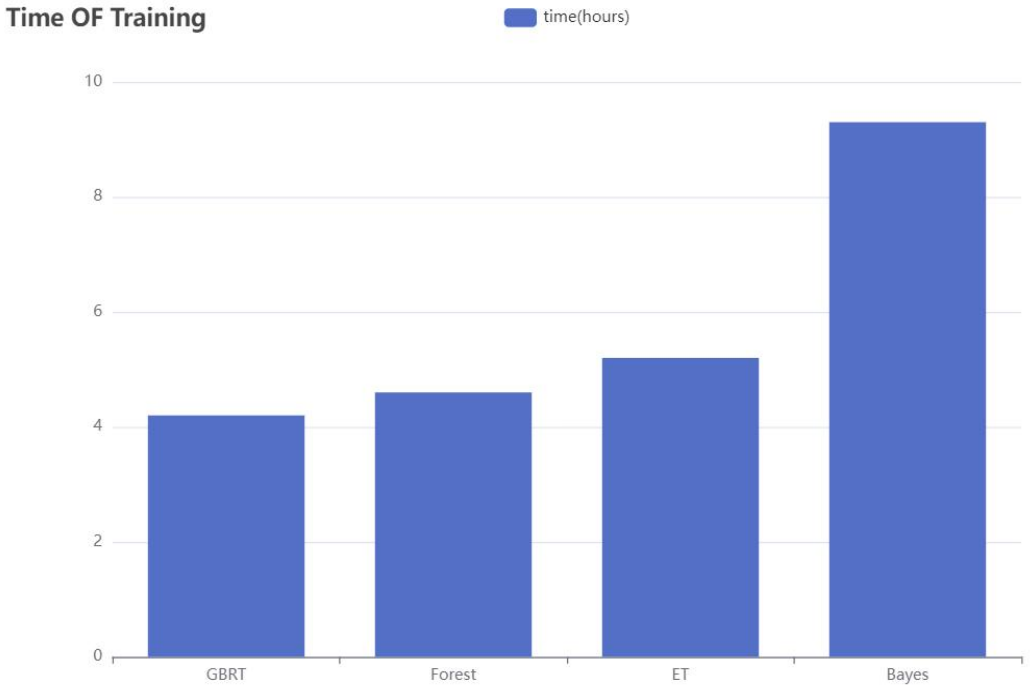
实验结果分析

4K顺序读

iops提升倍数：GBRT（9.3倍）、forest（8.4倍）、ET（9.9倍）、Bayes（5.8倍）

训练时长：GBRT（4.1小时）、forest（4.8小时）、ET（5.7小时）、Bayes（9.2小时）





实验结果分析

01

在4种算法中，ceph场景中，ET算法从优化效果上是最好的

02

贝叶斯算法优化效果不佳，并且耗时较长，尤其到了迭代周期的后半段

03

调优算法的效果好坏会根据场景不同而定

04

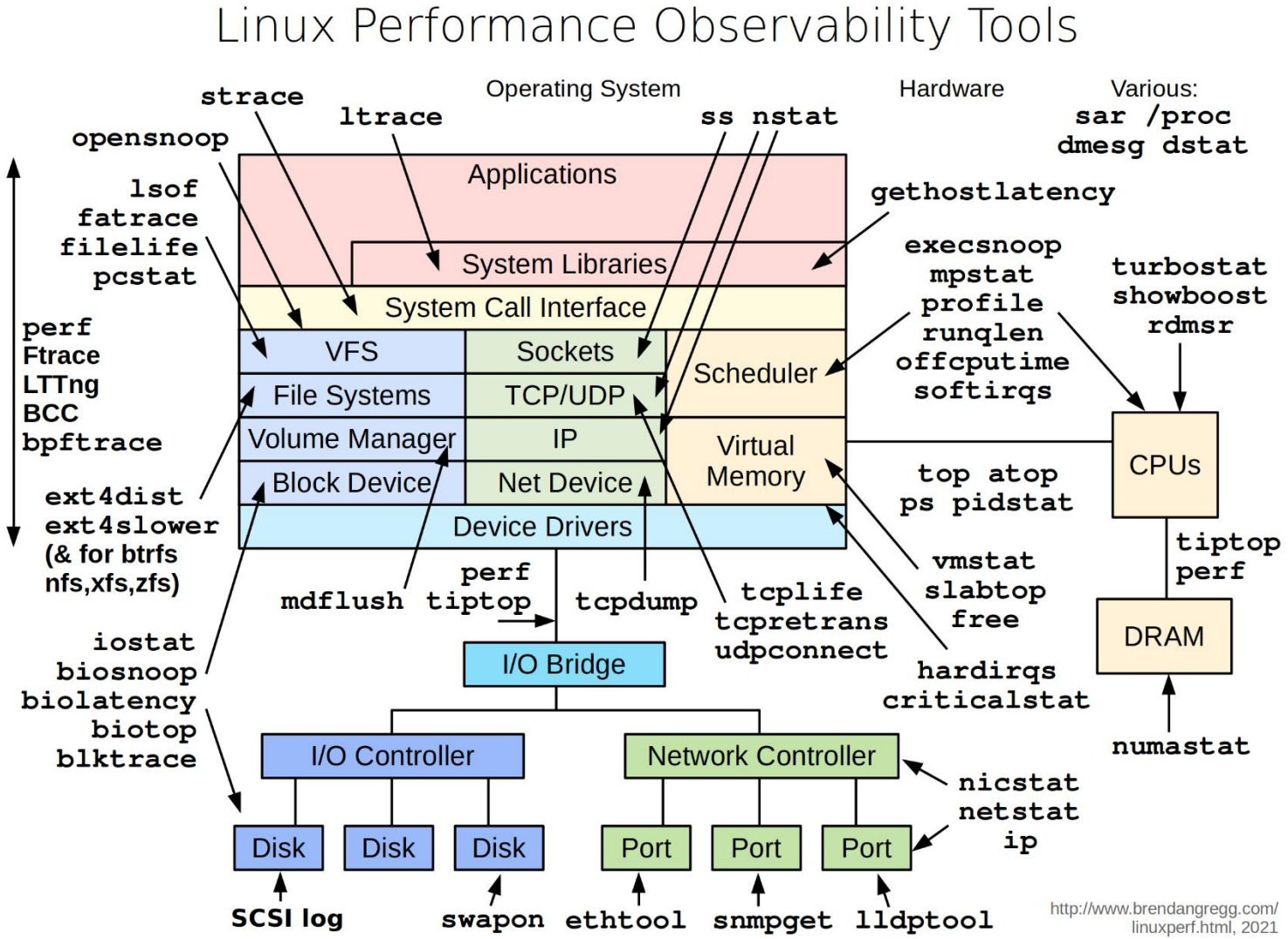
调优时间的长短根据配置参数的个数来决定

结 论

KYLINSOFT 麒麟软件 OpenEuler

延展

性能分析工具


Linux Performance Observability Tools
Operating System
Hardware
Various:
sar /proc
dmesg dstat
strace
ltrace
ss nstat
opensnoop
lsof
fatrace
filelife
pcstat
perf
Ftrace
LTTng
BCC
bpftrace
Applications
System Libraries
System Call Interface
VFS
Sockets
File Systems
TCP/UDP
Volume Manager
IP
Block Device
Net Device
Device Drivers
Scheduler
Virtual Memory
gethostlatency
execsnoop
mpstat
profile
runqlen
offcputime
softirqs
turbostat
showboost
rdmsr
CPUs
top atop
ps pidstat
tiptop
perf
DRAM
numastat
vmstat
slabtop
free
hardirqs
criticalstat
ext4dist
ext4slower
(& for btrfs
nfs,xfs,zfs)
mdflush
perf
tiptop
tcpdump
tcplife
tcpretrans
udpconnect
iostat
biosnoop
biolatency
biotop
blktrace
I/O Bridge
I/O Controller
Network Controller
nicstat
netstat
ip
Disk
Disk
Disk
Port
Port
Port
SCSI log
swapon
ethtool
snmpget
lldptool
http://www.brendangregg.com/
linuxperf.html, 2021

Atune发展

01

性能调优可视化

02

系统差异对比

03

全自动场景识别

THANKS

openEuler

KYLINSOFT
麒麟软件